CRISP-DM

Kelompok 1

Nama Kelompok:

- Afradillah
- Alvin
- Bahzar
- Dea
- Doni

Tools:

Table of Contents

Business Understanding

Jenis Bisnis

Perusahaan yang bergerak di bidang pengiriman barang atau produk yang biasa disebut sebagai perusahaan ekspedisi.

Divisi Customer Service

Customer care call, customer rating

Divisi Marketing

Reached on Time, Discount, Prior Purchase

Divisi Operasional

Warehouse Block, Mode of Shipment

Alasan

Customer rating yang fluktuatif.

Hipotesis

Apabila diskonnya semakin besar serta waktu pengirimannya semakin on time, customer baik laki-laki atau perempuan akan memberikan rating yang tinggi.

Business Objectives

Analisis pengaruh gender terhadap penilaian pelanggan berdasarkan diskon dan ketepatan waktu barang.

Tujuan Proyek

1. Mengetahui pengaruh nilai discount terhadap tingkat customer rating perusahaan ekspedisi dari customer laki-laki maupun perempuan.
2. Mengetahui pengaruh ketepatan waktu pengiriman terhadap tingkat customer rating perusahaan ekspedisi dari customer laki-laki maupun perempuan.

Data Understanding

Name of each column

```
df.columns
```

```
Index(['ID', 'Warehouse_block', 'Mode_of_Shipment', 'Customer_care_calls',
       'Customer_rating', 'Cost_of_the_Product', 'Prior_purchases',
       'Product_importance', 'Gender', 'Discount_offered', 'Weight_in_gms',
       'Reached.on.Time_Y.N'],
      dtype='object')
```

```
df.shape
```

```
(11005, 12)
```

```
df.size
```

```
132060
```

Shape and Size

Data Explorasi

Data Type

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11005 entries, 0 to 11004
Data columns (total 12 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   ID                   11005 non-null  int64
 1   Warehouse_block      11005 non-null  object
 2   Mode_of_Shipment     11005 non-null  object
 3   Customer_care_calls  10983 non-null  float64
 4   Customer_rating      11005 non-null  int64
 5   Cost_of_the_Product  11005 non-null  int64
 6   Prior_purchases      11005 non-null  int64
 7   Product_importance   11005 non-null  object
 8   Gender               11005 non-null  object
 9   Discount_offered     10905 non-null  float64
 10  Weight_in_gms        11005 non-null  int64
 11  Reached.on.Time_Y.N  11005 non-null  int64
dtypes: float64(2), int64(6), object(4)
memory usage: 1.0+ MB
```

Data Describe

	ID	Customer_care_calls	Customer_rating	Cost_of_the_Product	Prior_purchases	Discount_offered	Weight_in_gms	Reached.on.Time_Y.N
count	11005.000000	10983.000000	11005.000000	1.100500e+04	11005.000000	10905.000000	11005.000000	11005.000000
mean	5497.036801	4.052354	2.991095	9.296761e+03	3.567015	13.205777	3633.337937	0.598819
std	3176.951036	1.140826	1.413607	3.013132e+05	1.522694	16.084945	1635.315267	0.494038
min	1.000000	2.000000	1.000000	9.600000e+01	2.000000	1.000000	1001.000000	0.000000
25%	2746.000000	3.000000	2.000000	1.690000e+02	3.000000	4.000000	1839.000000	0.000000
50%	5497.000000	4.000000	3.000000	2.140000e+02	3.000000	7.000000	4148.000000	1.000000
75%	8248.000000	5.000000	4.000000	2.510000e+02	4.000000	10.000000	5049.000000	1.000000
max	10999.000000	7.000000	5.000000	1.000000e+07	10.000000	65.000000	7846.000000	3.000000

	Warehouse_block	Mode_of_Shipment	Product_importance	Gender
count	11005	11005	11005	11005
unique	6	3	3	4
top	F	Ship	low	F
freq	3661	7468	5297	5546

Data describe include object

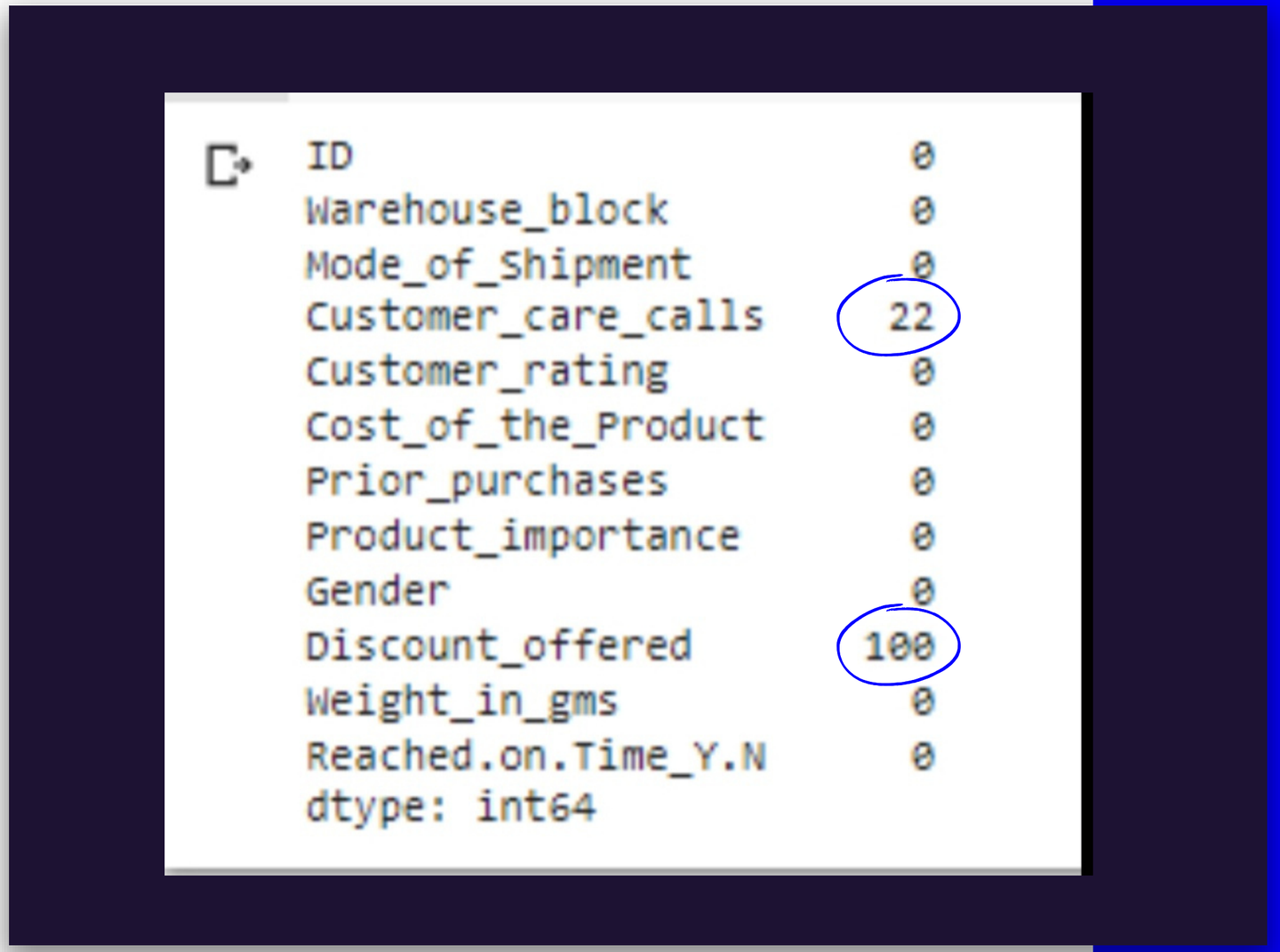
Checking Missing Values

- *Customer_care_calls*
- *Discount_offered*

Duplicate Row

df.duplicated().sum()

6



Unique Value

```
print(df.Warehouse_block.unique())
```

```
['D' 'F' 'A' 'B' 'C' 'ZX']
```

```
print(df.Mode_of_Shipment.unique())
```

```
['Flight' 'Ship' 'Road']
```

```
print(df.Customer_care_calls.unique())
```

```
[ 4.  2.  3.  5.  6.  7. nan]
```

```
print(df.Customer_rating.unique())
```

```
[2 5 3 1 4]
```

```
print(df.Cost_of_the_Product.unique())
```

```
[ 177  216  183  176  184  162  250  233
  150  164  189  232  198  275  152  227
  143  239  145  161  156  211  251  225
  172  234  266  257  223  149  137  181
  215  269  139  174  151  210  169  160
  190  141  165  170  203  246  238  193
  221  179  105  261  202  109  158  231
  206  187  230  113  180  132  217  197
  185  278  229  186  286  175  219  213
  235  271  144  218  263  168  265  205
  252  222  220  147  200  224  247  280
  157  207  237  264  248  191  146  135
   98   97  114  112  274  166  148  270
  242  192  116  255  209  134  130  133
  140  136  142  154  155  127  129  159
  294  226  258  241  208  182  115  212
  171  249  243  163  272  138  273  279
  173  194  262  201  260  188  267  131
  122  103  199  236  167  259  178  123
  124   96  244  254  128  204  245  228
  268  108  276  214  281  253  104  240
  121  153  111  117  195  110  119  196
  291  118  283  100  256  285  284  101
  296  277  106  282  126  102  120   99
  125  107  301  290  310  308  300  303
  306  292  293  295  304  298  305  287
  309  302  307  289  297  299  288 10000000]
```

```
print(df.Prior_purchases.unique())
```

```
[ 3  2  4  6  5  7 10  8]
```

```
print(df.Product_importance.unique())
```

```
['low' 'medium' 'high']
```

```
print(df.Gender.unique())
```

```
['F' 'M' 'Male' 'Female']
```

```
print(df.Discount_offered.unique())
```

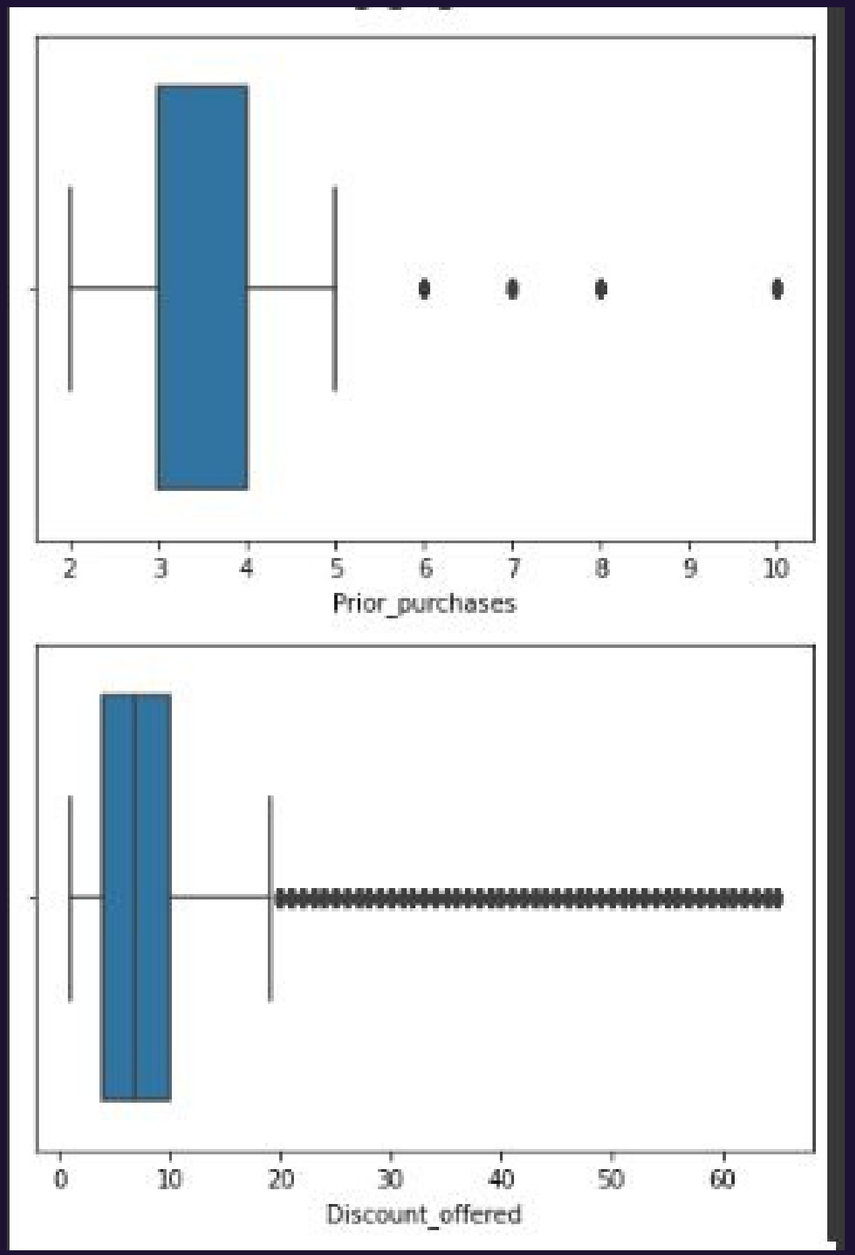
```
[44. 59. 48. 10. 46. 12.  3. 11. 29. 32.  1. 43. 45.  6. 36. 18. 38. 51.
  2. 28. 24. 31. 61. 22.  4. 62. 16. 56. 15.  9. 40. 37. 41. 17. 64. 52.
 49. 39. 14. 33. 21. 13. 23. 26. 57.  7. 35.  8.  5. 53. 55. 47. 65. 25.
 50. 60. 20. 19. 63. 58. 34. 54. 27. nan 30. 42.]
```

```
print(df.Weight_in_gms.unique())
```

```
[1233 3088 3374 ... 1086 1649 1098]
```

```
print(df['Reached.on.Time_Y.N'].unique())
```

```
[1 0 3]
```


2
3
4
5
6
7
8
9
10
Prior_purchases
0
10
20
30
40
50
60
Discount_offered

Outlier

Noise

Kolom Warehouse_block hanya berisi A,B,C,D,E, sehingga data F dan ZX adalah noise.

Warehouse_block

```
print(df.Warehouse_block.unique())
```

```
['D' 'F' 'A' 'B' 'C' 'ZX']
```

	ID	Warehouse_block	Mode_of_Shipment	Customer_care_calls	Customer_rating	Cost_of_the_Product	Prior_purchases	Product_importance	Gender	Discount_offered	Weight_in_gms	Reached.on.Time_Y.N
11000	10995	A	Ship	4.0	1	10000000	5	medium	F	1.0	1538	1
11001	10996	B	Ship	4.0	1	10000000	5	medium	F	6.0	1247	0
11002	10997	C	Ship	5.0	4	10000000	5	low	F	4.0	1155	0
11003	10998	F	Ship	5.0	2	10000000	6	medium	M	2.0	1210	0
11004	10999	D	Ship	2.0	5	10000000	5	low	F	6.0	1639	0

	ID	Warehouse_block	Mode_of_Shipment	Customer_care_calls	Customer_rating	Cost_of_the_Product	Prior_purchases	Product_importance	Gender	Discount_offered	Weight_in_gms	Reached.on.Time_Y.N
8766	8761	ZX	Road	4.0	3	187	6	low	F	9.0	5872	0
8767	8762	ZX	Road	3.0	4	136	3	medium	M	6.0	4631	0
8768	8763	ZX	Road	3.0	4	232	5	medium	M	9.0	5759	0
8769	8764	ZX	Flight	3.0	2	257	2	low	F	9.0	5085	1
8770	8765	ZX	Flight	4.0	1	156	2	medium	M	4.0	5225	0

Noise

Data 10000000 adalah noise di kolom Cost_of_the_Product

Cost_of_the_Product

```
177  216  183  176  184  162  250  233
150  164  189  232  198  275  152  227
143  239  145  161  156  211  251  225
172  234  266  257  223  149  137  181
215  269  139  174  151  210  169  160
190  141  165  170  203  246  238  193
221  179  105  261  202  109  158  231
206  187  230  113  180  132  217  197
185  278  229  186  286  175  219  213
235  271  144  218  263  168  265  205
252  222  220  147  200  224  247  280
157  207  237  264  248  191  146  135
 98   97  114  112  274  166  148  270
242  192  116  255  209  134  130  133
140  136  142  154  155  127  129  159
294  226  258  241  208  182  115  212
171  249  243  163  272  138  273  279
173  194  262  201  260  188  267  131
122  103  199  236  167  259  178  123
124   96  244  254  128  204  245  228
268  108  276  214  281  253  104  240
121  153  111  117  195  110  119  196
291  118  283  100  256  285  284  101
296  277  106  282  126  102  120   99
125  107  301  290  310  308  300  303
306  292  293  295  304  298  305  287
309  302  307  289  297  299  288 10000000]
```

ID	Warehouse_block	Mode_of_Shipment	Customer_care_calls	Customer_rating	Cost_of_the_Product	Prior_purchases	Product_importance	Gender	Discount_offered	Weight_in_gms	Reached.on.Time_Y.N
10995	A	Ship	4.0	1	10000000	5	medium	F	1.0	1538	1
10996	B	Ship	4.0	1	10000000	5	medium	F	6.0	1247	0
10997	C	Ship	5.0	4	10000000	5	low	F	4.0	1155	0
10998	F	Ship	5.0	2	10000000	6	medium	M	2.0	1210	0
10999	D	Ship	2.0	5	10000000	5	low	F	6.0	1639	0

Noise

Kolom Reached.on.Time_Y.N hanya berisi 1 dan 0, sehingga data 3 adalah noise.

Reached.on.Time_Y.N

```
print(df['Reached.on.Time_Y.N'].unique())
```

```
[1 0 3]
```

	ID	Warehouse_block	Mode_of_Shipment	Customer_care_calls	Customer_rating	Cost_of_the_Product	Prior_purchases	Product_importance	Gender	Discount_offered	Weight_in_gms	Reached.on.Time_Y.N
8257	8252	F	Ship	5.0	3	226	2	high	F	6.0	5033	3
8258	8253	A	Ship	5.0	2	158	3	low	M	4.0	4324	3
8259	8254	B	Ship	3.0	2	274	2	high	M	2.0	5029	3
8260	8255	C	Ship	3.0	4	245	3	medium	F	7.0	5429	3
8261	8256	F	Ship	5.0	3	254	6	low	M	8.0	1932	3

Inkonsistensi

Data dalam kolom Gender menunjukkan inkonsistensi

```
['F' 'M' 'Male' 'Female']
```

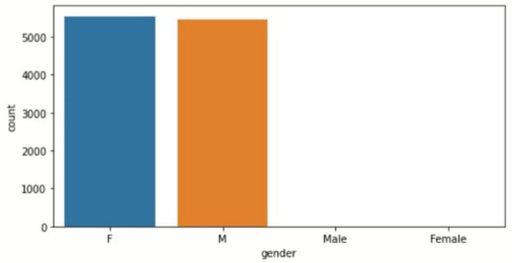
	ID	Warehouse_block	Mode_of_Shipment	Customer_care_calls	Customer_rating	Cost_of_the_Product	Prior_purchases	Product_importance	Gender	Discount_offered	Weight_in_gms	Reached.on.Time_Y.N
295	290	F	Flight	4.0	2	172	3	medium	Female	59.0	1858	1
296	291	A	Flight	4.0	4	270	3	low	M	12.0	3279	1
297	292	B	Ship	4.0	1	206	3	medium	M	23.0	3807	1
298	293	C	Ship	4.0	3	211	4	medium	M	40.0	3106	1
299	294	F	Ship	3.0	2	136	4	low	M	62.0	1090	1
300	295	D	Ship	5.0	1	186	2	high	F	52.0	3981	1
301	296	F	Ship	3.0	1	202	2	medium	Male	62.0	3398	1
302	297	A	Ship	4.0	2	183	3	medium	F	12.0	3646	1
303	298	B	Ship	4.0	4	232	2	low	M	36.0	2861	1
304	299	C	Ship	4.0	2	147	4	medium	M	52.0	1129	1
305	300	F	Ship	4.0	3	136	3	medium	M	20.0	1531	1
306	301	D	Ship	5.0	3	127	3	low	Female	43.0	1446	1
307	302	F	Ship	4.0	5	202	2	medium	M	35.0	3695	1

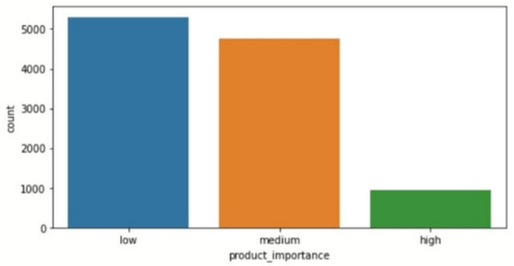
Check String

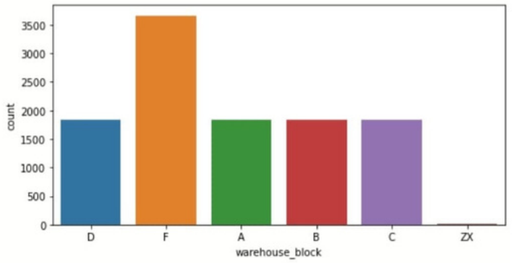
Nama kolom terlihat tidak rapih

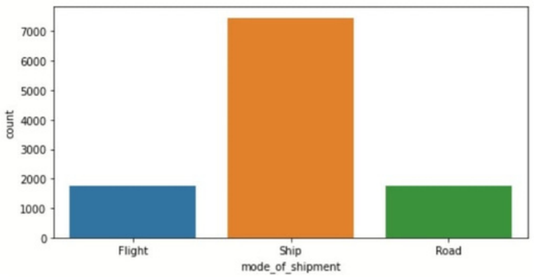
	ID	Warehouse_block	Mode_of_Shipment	Customer_care_calls	Customer_rating	Cost_of_the_Product	Prior_purchases	Product_importance	Gender	Discount_offered	Weight_in_gms	Reached.on.Time_Y_N
8766	8761	ZX	Road	4.0	3	187	6	low	F	9.0	5872	0
8767	8762	ZX	Road	3.0	4	136	3	medium	M	6.0	4631	0
8768	8763	ZX	Road	3.0	4	232	5	medium	M	9.0	5759	0
8769	8764	ZX	Flight	3.0	2	257	2	low	F	9.0	5085	1
8770	8765	ZX	Flight	4.0	1	156	2	medium	M	4.0	5225	0

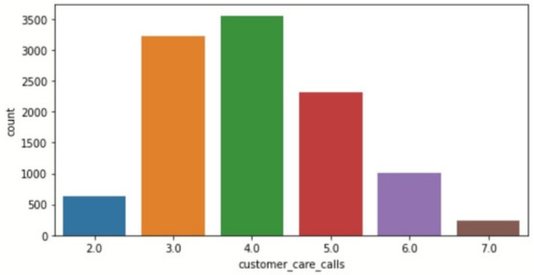
Initial EDA

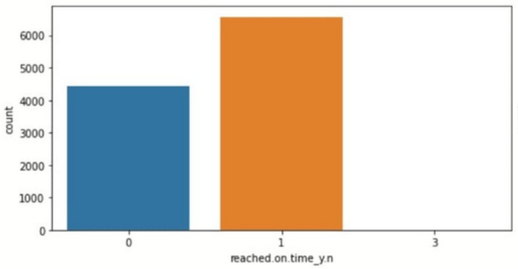


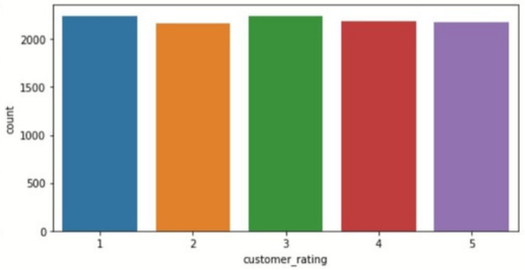












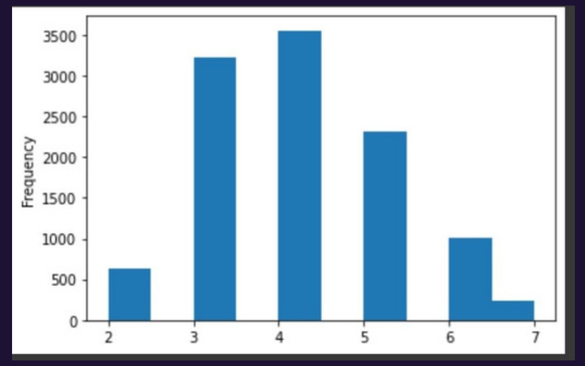
- Jumlah Female 5546 + 5 (F), Male 5447 + 7(M).
- Para customer, lebih sering menggunakan jasa perusahaan untuk mengirimkan product dengan kepentingan yang rendah.
- Produk-produk milik customer, lebih banyak disimpan di gudang besar F.
- Customer lebih memilih kapal menjadi mode pengiriman produknya.
- Kebanyakan customer sering menghubungi shipper sebanyak 4 kali.

- Ada lebih dari 4000 product yang tidak mengalami keterlambatan, dan 6500 lebih yang mengalami keterlambatan. Kemudian ada sedikit noise pada kolom Reached.On.Time_Y.N karena (3) tidak sesuai dengan data dictionary.
- Kolom customer_rating memiliki data yang bervariasi, masing-masing rating selalu di atas 2000.
- Rata-rata terbanyak untuk diskon yang ditawarkan adalah berkisar antara 0-15% untuk tiap product.

Data Preparation

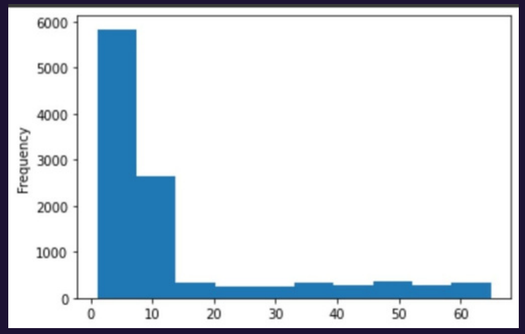
Handling Missing Value

Customer Care Call



Sebarannya adalah normal, sehingga diisi dengan mean

Discount Offered



Sebarannya adalah skew, sehingga diisi dengan median

```
ID                    0
Warehouse_block       0
Mode_of_Shipment      0
Customer_care_calls   0
Customer_rating       0
Cost_of_the_Product   0
Prior_purchases       0
Product_importance    0
Gender                0
Discount_offered      0
Weight_in_gms         0
Reached.on.Time_Y.N   0
dtype: int64
```

After Filling

Handling Duplicate

Duplicate rows harus dihapus karena rowsnya berisi value yang sama sehingga akan mempengaruhi jumlah data

Before

	ID	Warehouse_block	Mode_of_Shipment	Customer_care_calls	Customer_rating	Cost_of_the_Product	Prior_purchases	Product_importance	Gender	Discount_offer
21	21	A	Ship	3.0	3	161	2	medium	F	3
22	21	A	Ship	3.0	3	161	2	medium	F	3
23	21	A	Ship	3.0	3	161	2	medium	F	3
112	109	D	Ship	4.0	5	238	3	high	F	5
113	109	D	Ship	4.0	5	238	3	high	F	5
114	109	D	Ship	4.0	5	238	3	high	F	5

After

```
#check for duplicated rows
df.duplicated().sum()
```

```
0
```

```
df[df[['ID']].duplicated()]
```

ID	Warehouse_block	Mode_of_Shipment	Customer_care_calls	Customer_rating	Cost_of_the_Product	Prior_purchases	Product_importance	Gender	Discount_offered

Handling Noise

Warehouse_block

- Replace F to E
- Drop ZX

```
print(df.Warehouse_block.unique())
```

```
['D' 'E' 'A' 'B' 'C']
```

Cost_of_the_product

```
df['cost_of_the_product'].max()
```

```
310
```

Drop 10000000

Reached.On.Time_Y.N

```
print(df['Reached.on.Time_Y.N'].unique())
```

```
[1 0]
```

Drop 3

Handling Inconsistency

Gender

Replace F to Female
and M to male

```
print(df.gender.unique())
```

```
['Female' 'Male']
```

Data F dan M di replace agar data yang dihasilkan lebih konsisten

Column rename

Nama column di rename ke lowercase agar lebih rapih

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11005 entries, 0 to 11004
Data columns (total 12 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   id                   11005 non-null  int64
 1   warehouse_block      11005 non-null  object
 2   mode_of_shipment     11005 non-null  object
 3   customer_care_calls  11005 non-null  int64
 4   customer_rating      11005 non-null  int64
 5   cost_of_the_product  11005 non-null  int64
 6   prior_purchases      11005 non-null  int64
 7   product_importance   11005 non-null  object
 8   gender               11005 non-null  object
 9   discount_offered     11005 non-null  float64
 10  weight_in_gms        11005 non-null  int64
 11  reached.on.time_y.n  11005 non-null  int64
dtypes: float64(1), int64(7), object(4)
memory usage: 1.0+ MB
```

Data type correction

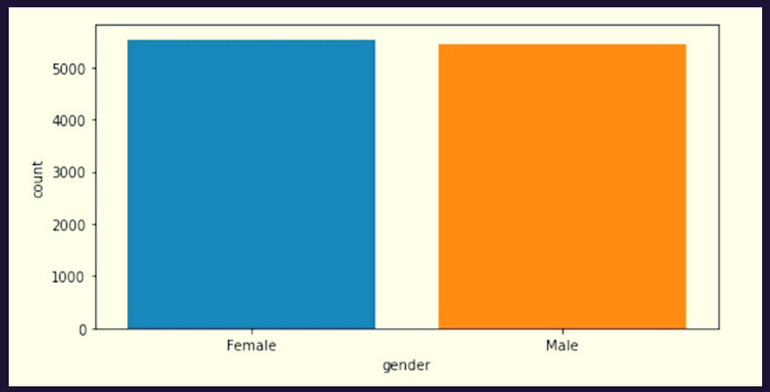
Mengubah data type customer care calls dari "Float" ke "Integer", karena jumlah panggilan seharusnya tidak berbentuk decimal

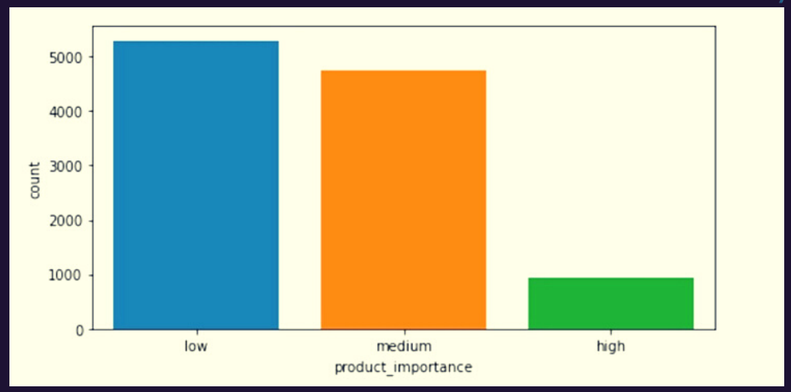
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11005 entries, 0 to 11004
Data columns (total 12 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   ID                   11005 non-null  int64
 1   Warehouse_block      11005 non-null  object
 2   Mode_of_Shipment     11005 non-null  object
 3   Customer_care_calls  11005 non-null  int64
 4   Customer_rating      11005 non-null  int64
 5   Cost_of_the_Product  11005 non-null  int64
 6   Prior_purchases      11005 non-null  int64
 7   Product_importance   11005 non-null  object
 8   Gender               11005 non-null  object
 9   Discount_offered     11005 non-null  float64
 10  Weight_in_gms        11005 non-null  int64
 11  Reached.on.Time_Y.N  11005 non-null  int64
dtypes: float64(1), int64(7), object(4)
memory usage: 1.0+ MB
```

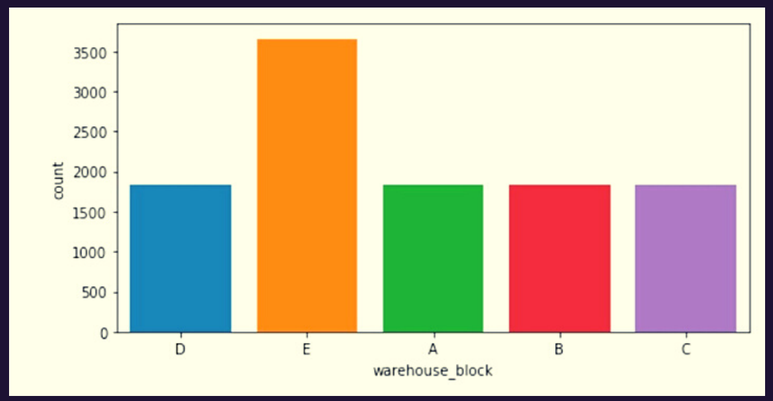
How about outlier?

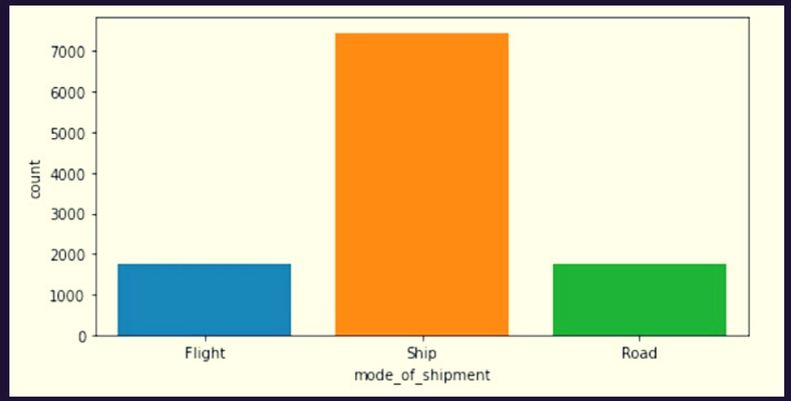
Outlier tidak di drop karena sekecil apapun datanya, dapat menjadi bahan analisis perusahaan. Sehingga dengan analisis tersebut harapannya dapat menjadi bahan evaluasi perusahaan.

Modeling (EDA)



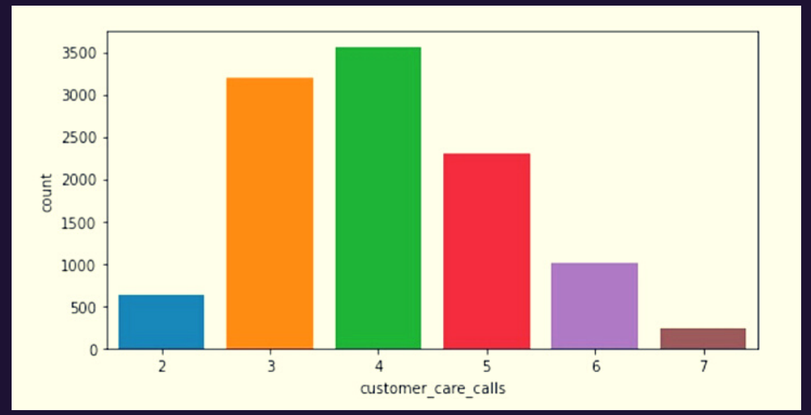


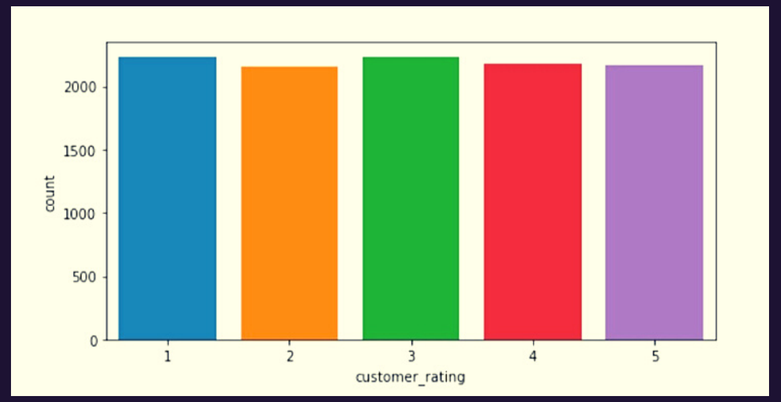


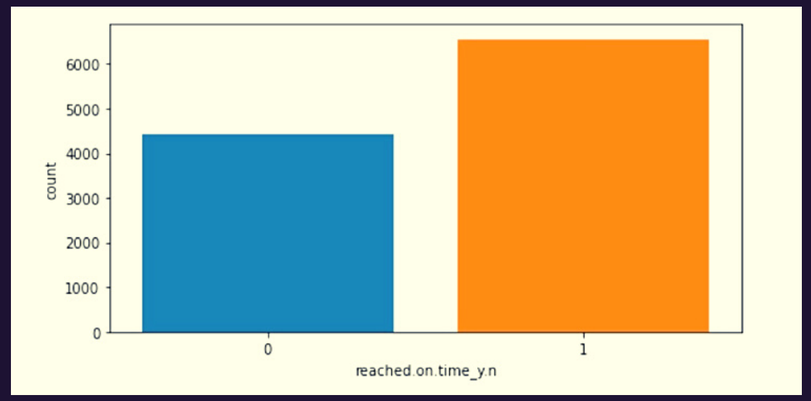


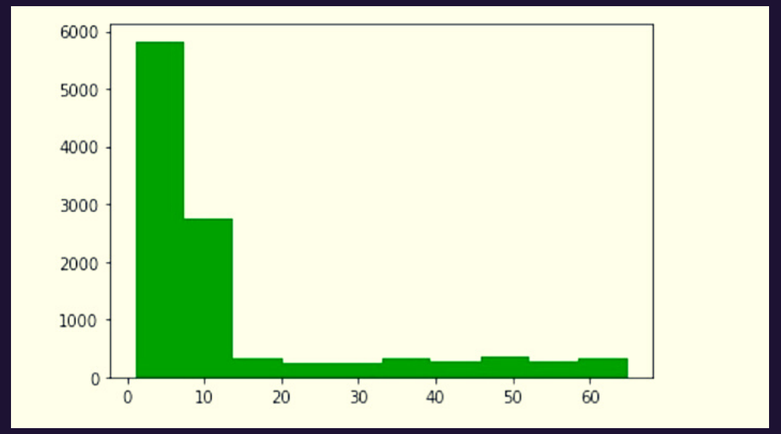
Notes:

- Jumlah customer perempuan sebanyak 5534 dan laki-laki sebanyak 5443.
- Customer lebih sering menggunakan jasa perusahaan untuk mengirim produk berkepentingan rendah (low).
- Produk customer lebih banyak disimpan di warehouse E.
- Customer lebih memilih kapal sebagai moda pengiriman produk, sehingga besar kemungkinan daerah pengiriman adalah antar pulau atau bahkan antar negara.





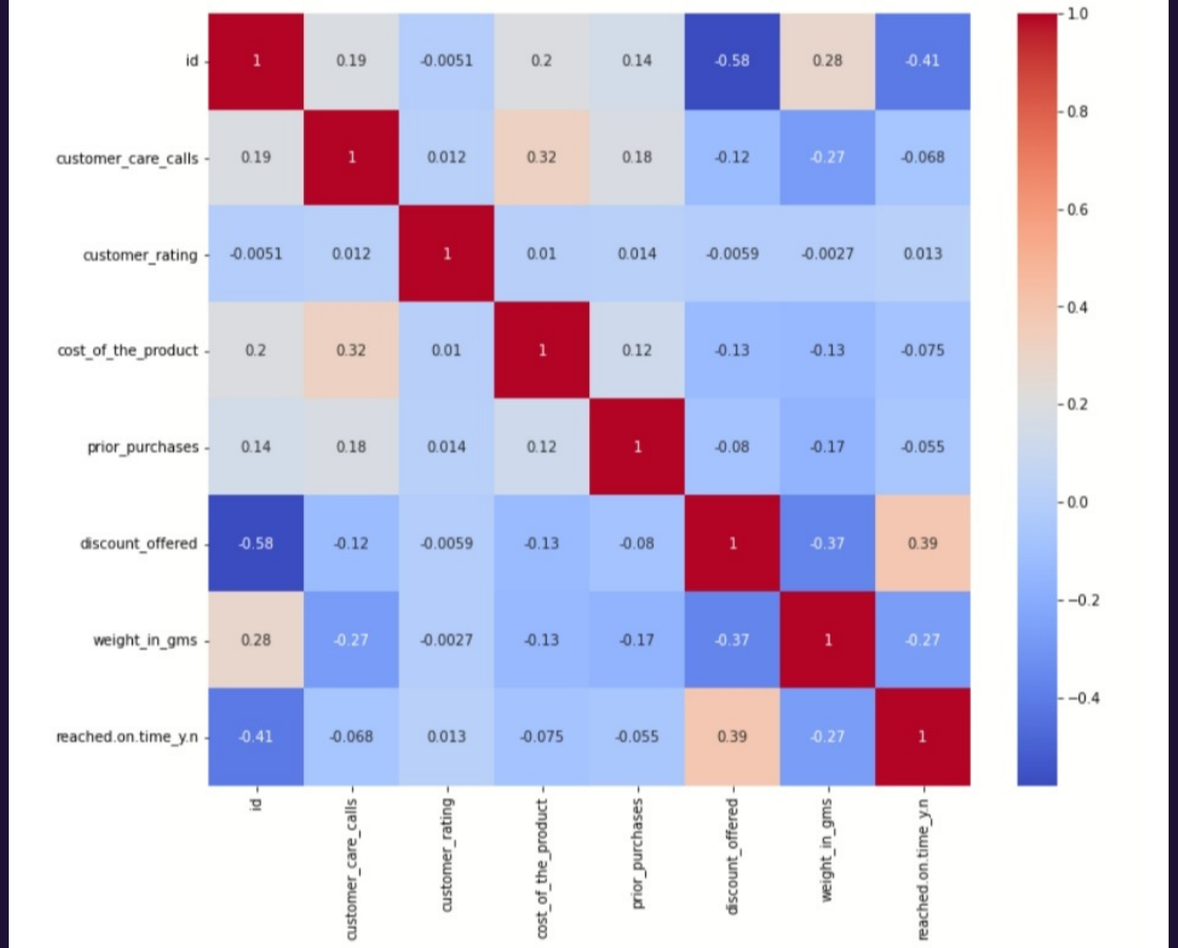




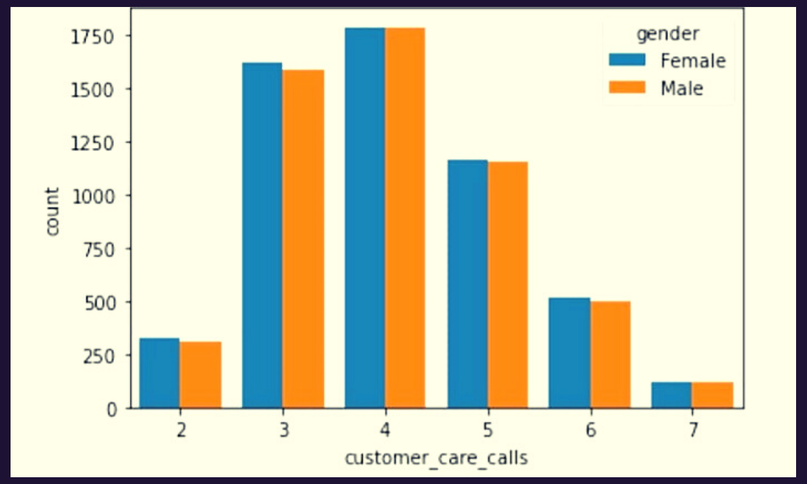
Persebaran distribusi pemberian diskon

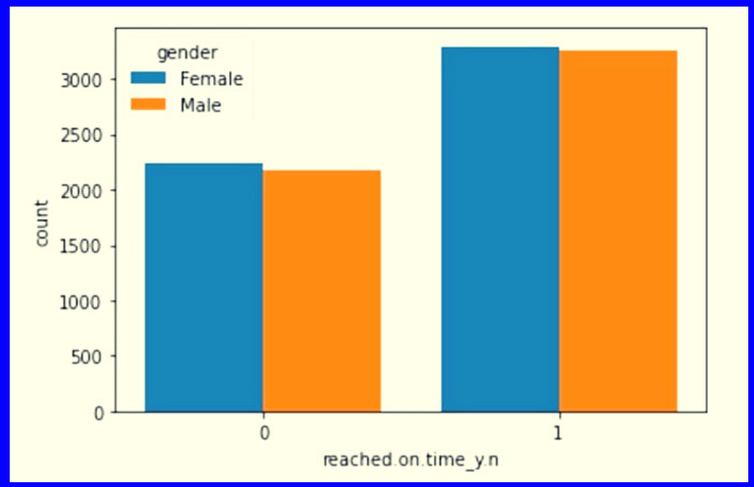
Notes:

- Secara umum, kebanyakan customer menghubungi shipper sebanyak empat kali.
- Ada lebih dari 4000 product yang on time dan ada 6000 lebih yang terlambat. Artinya jumlah product yang terlambat lebih besar daripada yang on time.
- Rating sangatlah bervariasi dan masing-masing rating jumlahnya selalu di atas 2000.
- Diskon yang ditawarkan rata-rata terbanyak adalah berkisar pada 0-15% untuk tiap product.



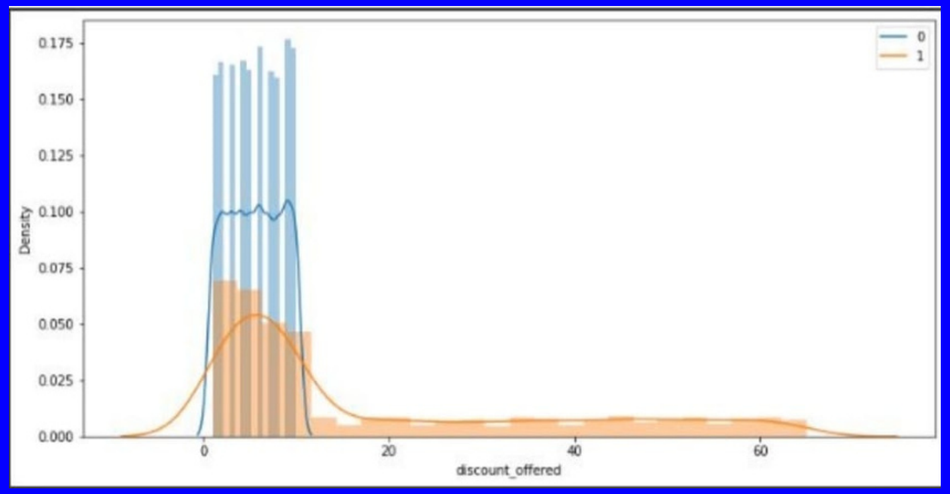
Korelasi antar variabel

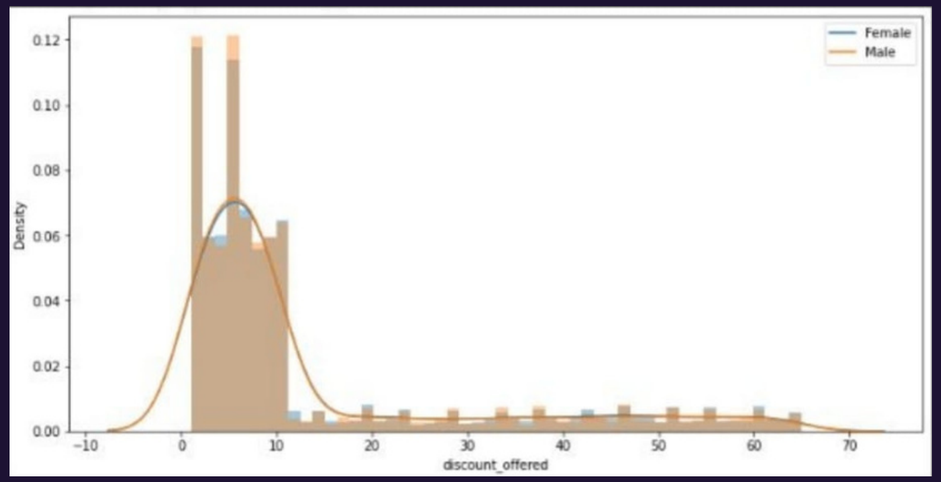




Berdasarkan data tersebut, perempuan lebih sering menghubungi customer call pada saat pengiriman barangnya, jadi kemungkinan ada hal-hal yang tidak berjalan dengan semestinya atau terdapat permasalahan pada pengiriman.

Hipotesis pada visualisasi sebelumnya yang menyatakan bahwa terdapat masalah keterlambatan adalah benar, karena customer perempuan mendapatkan jumlah pengiriman yang terlambat lebih banyak daripada laki-laki.



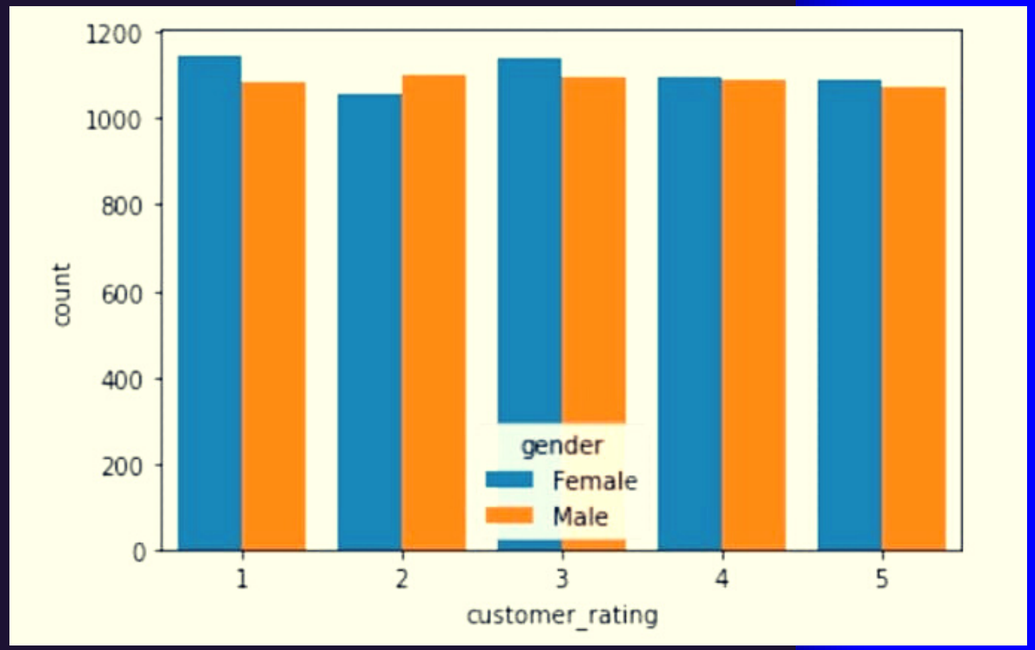


Berdasarkan visualisasi data tersebut, terlihat bahwa apabila diskon yang diberikan kepada customer semakin tinggi, maka cenderung akan mengalami keterlambatan pengiriman. Sehingga asumsinya adalah bahwa terdapat prioritas dalam customer yang mendapatkan diskon kecil (1-10%).

Dalam visualisasi tersebut terlihat bahwa customer laki-laki yang merupakan penerima diskon rendah (1-10%), artinya benar adanya bahwa customer perempuan sering menghubungi shipper karena adanya keterlambatan pengiriman.

Kesimpulannya,

- Yang banyak mendapat diskon rendah (1-10%) adalah laki-laki. Artinya berdasarkan korelasi antara diskon dengan ketepatan waktu pengiriman maka diskon 1-10% cenderung rata-rata tidak mengalami keterlambatan.
- Cust perempuan lebih sering menghubungi shipper karena masalah keterlambatan.
- Hasil korelasi antara diskon dan ketepatan waktu pengiriman adalah semakin tinggi diskon, semakin tinggi juga keterlambatan. Karena nilai korelasi 0,39 arah geraknya positif (berbanding lurus).



CONT...

CONT....

- Hubungan antara diskon yang ditawarkan dengan rating yang diberikan oleh customer, analisanya adalah discount tidak langsung mempengaruhi terhadap cust rating, karena jika dilihat dari korelasi, hanya sebesar -0,005. Kemudian, diambil insight lagi pada korelasi antara discount dengan reached on time, 0.39 dengan arah positif, artinya semakin tinggi diskon maka semakin tinggi pula keterlambatan pengirimannya. Jadi, karena ada permasalahan keterlambatan pengiriman, yang salah satunya jika dilihat dari jalan cerita analisis EDA nya adalah diakibatkan dari variabel diskon, maka kami simpulkan diskon memengaruhi pemberian rating oleh customer. Korelasinya -0,0059 antara diskon dengan cust rating dengan arah negatif, semakin besar diskon maka cust rating semakin rendah.

Berdasarkan story pada EDA tersebut, maka sudah diketahui penyebab customer perempuan lebih banyak memberi rating 1-3 (rendah), yakni adanya permasalahan keterlambatan pengiriman.

Rekomendasi

Meningkatkan profesionalitas perusahaan dalam melakukan pelayanan terhadap customer.

Perusahaan memberikan prioritas yang sama dalam pengiriman sehingga pengiriman dapat tepat waktu, karena hal tersebut berpengaruh terhadap kepuasan konsumen

Mengevaluasi pemberian diskon dengan menyesuaikan proporsinya seperti hanya diberikan di hari-hari atau momen tertentu.